

Algorithms in the Ultra-Wide Word Model

Arash Farzan¹, Alejandro López-Ortiz², Patrick K. Nicholson³, and Alejandro Salinger⁴

¹ Facebook Inc.
afarzan@fb.com

² David R. Cheriton School of Computer Science, University of Waterloo
alopez-o@uwaterloo.ca

³ Max-Planck-Institut für Informatik
pnichols@mpi-inf.mpg.de

⁴ Department of Computer Science, Saarland University
salinger@cs.uni-saarland.de

Abstract. The effective use of parallel computing resources to speed up algorithms in current multi-core parallel architectures remains a difficult challenge, with ease of programming playing a key role in the eventual success of various parallel architectures. In this paper we consider an alternative view of parallelism in the form of an ultra-wide word processor. We introduce the Ultra-Wide Word architecture and model, an extension of the word-RAM model that allows for constant time operations on thousands of bits in parallel. Word parallelism as exploited by the word-RAM model does not suffer from the more difficult aspects of parallel programming, namely synchronization and concurrency. For the standard word-RAM algorithms, the speedups obtained are moderate, as they are limited by the word size. We argue that a large class of word-RAM algorithms can be implemented in the Ultra-Wide Word model, obtaining speedups comparable to multi-threaded computations while keeping the simplicity of programming of the sequential RAM model. We show that this is the case by describing implementations of Ultra-Wide Word algorithms for dynamic programming and string searching. In addition, we show that the Ultra-Wide Word model can be used to implement a non-standard memory architecture, which enables the sidestepping of lower bounds of important data structure problems such as priority queues and dynamic prefix sums. While similar ideas about operating on large words have been mentioned before in the context of multimedia processors [37], it is only recently that an architecture like the one we propose has become feasible and that details can be worked out.

1 Introduction

In the last few years, multi-core architectures have become the dominant commercial hardware platform. The potential of these architectures to improve performance through parallelism remains to be fully attained, as effectively using all cores on a single application has proven to be a difficult challenge. In this

paper we introduce the Ultra-Wide Word architecture and model of computation, an alternate view of parallelism for a modern architecture in the form of an ultra-wide word processor. This can be implemented by replacing one or more cores of a multi-core chip with a very wide word Arithmetic Logic Unit (ALU) that can perform operations on a very large number of bits in parallel.

The idea of executing operations on a large number of bits simultaneously has been successfully exploited in different forms. In Very Long Instruction Word (VLIW) architectures [17], several instructions can be encoded in one wide word and executed in one single parallel instruction. Vector processors allow the execution of one instruction on multiple elements simultaneously, implementing Single-Instruction-Multiple-Data (SIMD) parallelism. This form of parallelism led to the design of supercomputers such as the Cray architecture family [36] and is now present in Graphics Processing Units (GPUs) as well as in Streaming SIMD Extensions (SSE) to scalar processors.

In 2003, Thorup [37] observed that certain instructions present in some SSE implementations were particularly useful for operating on large integers and speeding up algorithms for combinatorial problems. To a certain extent, some of the ideas in the Ultra Wide Word architecture are presaged in the paper by Thorup, which was proposed in the context of multimedia processors. Our architecture developed independently and differs on several aspects (see discussion in Section 2.3) but it is motivated by similar considerations.

As CPU hardware advances, so does the model used in theory to analyze it. The increase in word size was reflected in the word-RAM model in which algorithm performance is given as a function of the input size n and the word size w , with the common assumption that $w = \Theta(\log n)$. In its simplest version, the word-RAM model allows the same operations as the traditional RAM model. Algorithms in this model take advantage of bit-level parallelism through packing various elements in one word and operating on them simultaneously. Although similar to vector processing, the word-RAM provides more flexibility in that the layout of data in a word depends on the algorithm and data elements can be packed in an arbitrary way. Unlike VLIW architectures, the Ultra-Wide Word model we propose is not concerned with the compiler identifying operations which can be done in parallel but rather with achieving large speedups in implementations of word-RAM algorithms through operations on thousands of bits in parallel.

As multi-core chip designs evolve, chip vendors try to determine the best way to use the available area on the chip, and the options traditionally are an increased number of cores or larger caches. We believe that the current stage in processor design allows for the inclusion of an architecture such as the one we propose. In addition, ease of programming is a major hurdle to the eventual success of parallel and multi-core architectures. In contrast, bit parallelism as exploited by the word-RAM model does not suffer from this drawback: there is a large selection of word-RAM algorithms (see, e.g., [2, 26, 24, 12]) that readily benefit from bit parallelism without having to deal with the more difficult aspects of concurrency such as mutual exclusion, synchronization, and resource contention. In this sense, the advantage of an on-chip ultra-wide word architecture is that

it can enable word-RAM algorithms to achieve speedups comparable to those of multi-threaded computations, while at the same time keeping the simplicity of sequential programming that is inherent to the RAM model. We argue that this is the case by showing several examples of implementations of word-RAM algorithms using the wide word, usually with simple modifications to existing algorithms, and extending the ideas and techniques from the word-RAM model.

In terms of the actual architecture, we envision the ultra-wide ALU together with multi-cores on the same chip. Thus, the Ultra-Wide Word architecture adds to the computing power of current architectures. The results we present in this paper, however, do not use multi-core parallelism.

Summary of Results We introduce the Ultra-Wide Word architecture and model, which extends the w -bit word-RAM model by adding an ALU that operates on w^2 -bit words. We show that several broad classes of algorithms can be implemented in this model. In particular:

- We describe Ultra-Wide Word implementations of dynamic programming algorithms for the subset sum problem, the knapsack problem, the longest common subsequence problem, as well as many generalizations of these problems. Each of these algorithms illustrates a different technique (or combination of techniques) for translating an implementation of an algorithm in the word-RAM model to the Ultra-Wide Word model. In all these cases we obtain a w -fold speedup over word-RAM algorithms.
- We also describe Ultra-Wide Word implementations of popular string searching algorithms: the Shift-And/Shift-Or algorithms [4, 40] and the Boyer-Moore-Horspool algorithm [28]. Again, we obtain a w -fold speedup over the original algorithms.
- Finally, we show that the Ultra-Wide Word model is powerful enough to simulate a non-standard memory architecture in which bytes can overlap, which we shall call FS-RAM [18]. This allows us to implement data structures and algorithms that circumvent known lower bounds for the word-RAM model.

The rest of this paper is organized as follows. In Section 2 we describe the Ultra-Wide architecture and model of computation. We show in Section 3 how to simulate the FS-RAM memory architecture. In Sections 4 and 5 we present UW-RAM implementations of algorithms for dynamic programming and string searching. We present concluding remarks in Section 6.

2 The Ultra-Wide Word-RAM Model

The Ultra-Wide word-RAM model (UW-RAM) we propose is an extension of the word-RAM model. We briefly review here the key features of the word-RAM.

2.1 Algorithms in the word-RAM model

The word-RAM is a variant of the RAM model in which a word has length w bits, and the contents of memory are integers in the range $\{0, \dots, 2^w - 1\}$ [24]. This implies that $w \geq \log n$, where n is the size of the input, and a common assumption is $w = \Theta(\log n)$ (see, e.g., [32, 8]). The word-RAM includes the usual load, store, and jump instructions of the RAM model, allowing for immediate operands and for direct and indirect addressing. In this model, arithmetic operations on two words are modulo 2^w , and the instruction set includes left and right shift operations (equal to multiplication and division by powers of two) and boolean operations. All instructions take constant time to execute. There are different versions of the word-RAM model depending on the instruction set assumed to be available. The *restricted model* is limited to addition, subtraction, left and right shifts, and boolean operations AND, OR, and NOT. These instructions augmented with multiplication constitute the *multiplication model*. Finally, the AC^0 model assumes that all functions computable by an unbounded fan-in circuit of polynomial size (in w) and constant depth are available in the instruction set and execute in constant time. This definition includes all instructions from the restricted model and excludes multiplication. We refer to the reader to the survey by Hagerup [24] for a more extended description of the model and a discussion of its practicality.

Word-RAM algorithms exploit word-level parallelism by operating on various elements simultaneously using instructions on w -bits words. There are various algorithms for fundamental problems that take advantage of word-level parallelism or a bounded universe, some of which fit into the word-RAM model, although are not explicitly designed for it [3]. Much attention has been given to sorting and searching, for which known lower bounds in the comparison model do not carry to the word-RAM model [20]. For example, in a word-RAM model with multiplication, sorting n words can be done in $O(n \log \log n)$ time and $O(n)$ space deterministically [26], and in expected $O(n \sqrt{\log \log n})$ time and $O(n)$ space using randomization [27]. Word-RAM techniques have also been applied in many different areas, such as succinct data structures [29, 32], computational geometry [12, 13], and text indexing [22].

2.2 Ultra-Wide RAM

The Ultra-Wide word-RAM model (UW-RAM) extends the word-RAM model by introducing an ultra-wide ALU with w^2 -bit *wide words*, where w is the number of bits in a word-RAM. The ultra-wide ALU supports the basic operations available in a word-RAM on the entire word at once. As in the word-RAM model, the available set of instructions can be assumed to be those of the restricted, multiplication, or the AC^0 models. For the results in this paper we assume the instructions of the restricted model (addition, subtraction, left and right shift, and bitwise boolean operations), plus two non-standard straightforward AC^0 operations that we describe at the end of this subsection.

The model maintains the standard w -bit ALU as well as w -bit memory addressing. In general, we use the parameter w for the word size in the description and analysis of algorithms, although in some cases we explicitly assume $w = \Theta(\log n)$. In terms of real world parameters, the wide word in the ultra-wide ALU would presently have between 1,000 and 10,000 bits and could increase even further in the future. In reality, the addition of an ALU that supports operations on thousands on bits would require appropriate adjustments to the data and instruction caches of a processor as well as to the instruction pipeline implementation. Similarly to the abstractions made by the RAM and word-RAM models, the UW-RAM model ignores the effects of these and other architectural features and assumes that the execution of instructions on ultra-wide words is as efficient as the execution of operations on regular w -bit words, up to constant factors.

Provided that the UW-RAM supports the same operations as the word-RAM, the techniques to achieve bit-level parallelism in the word-RAM extend directly to the UW-RAM. However, since the word-RAM assumes that a word can be read from memory in constant time, many operations in word-RAM algorithms can be implemented through constant time table lookups. For example, counting the number of set bits in a word of $w = \log n$ bits can be implemented through two table lookups to a precomputed table that stores the number of set bits for each number of $\log n/2$ bits. The space used by the table is \sqrt{n} words. We cannot expect to achieve the same constant time lookup operation with words of w^2 bits since the size of the lookup tables would be prohibitive. However, the memory access operations of our model allow for the implementation of simultaneous table lookups of several w -bit words within a wide word, as we shall explain below.

We first introduce some notation. Let W denote a w^2 -bit word. Let $W[i]$ denote the i -th bit of W , and let $W[i..j]$ denote the contiguous subword of W from bit i to bit j , inclusive. The least significant bit of W is $W[0]$, and thus $W = \sum_{i=0}^{w^2-1} W[i] \times 2^i$. For the sake of memory access operations, we divide W into w -bit blocks. Let W_j denote the j -th contiguous block of w bits in W , for $0 \leq j \leq w-1$, and let $W_j[i]$ denote the i -th bit within W_j . Thus, $W_j = W[jw..(j+1)w-1]$ and $W = \sum_{j=0}^{w-1} 2^{jw} \times (\sum_{i=0}^{w-1} W_j[i] \times 2^i)$. The division of a wide word in blocks is solely intended for certain memory access operations, but basic operations of the model have no notion of block boundaries. Fig. 1 shows a representation of a wide word, depicting bits with increasing significance from left to right. In the description of operations with wide words we generally refer to variables with uppercase letters, whereas we use lowercase to refer to regular variables that use one w -bit word. Thus, shifts to the left (right) by i are equivalent to division (multiplication) by 2^i . In addition, we use $\mathbf{0}$ to denote a wide word with value 0. We use standard C-like notation for operations AND ('&'), OR ('|'), NOT ('~') and shifts ('<<', '>>').

Memory Access Operations In this architecture w (not necessarily contiguous) words from memory can be transferred into the w blocks of a wide word W in constant time. These blocks can be written to memory in parallel as well. As

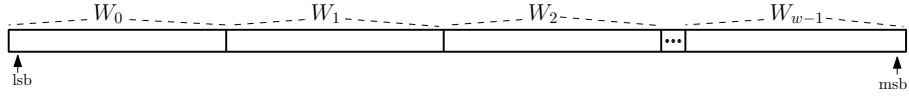


Fig. 1. A wide word in the Ultra-Wide Word architecture. The wide word is divided in w blocks of w bits each, shown here in increasing number of block from left to right.

with PRAM algorithms, the memory access type of the model can be assumed to allow or disallow concurrent reads and writes. For the results in this paper we assume the Concurrent-Read-Exclusive-Write (CREW) model.

The memory access operations that involve wide words are of three types: block, word, and content. We describe read accesses (write accesses are analogous). A *block access* loads a single w -bit word from memory into a given block of a wide word. A *word access* loads w contiguous w -bit words from memory into an entire wide word in constant time. Finally, a *content access* uses the contents of a wide word W as addresses to load (possibly non-contiguous) words of memory simultaneously: for each block j within W , this operation loads from memory the w -bit word whose address is W_j (plus possibly a base address). The specifics of read and write operations are shown in Table 1.

Note that accessing several (possibly non-contiguous) words from memory simultaneously is an assumption that is already made by any shared memory multiprocessing model. While, in reality, simultaneous access to all addresses in actual physical memory (e.g., DRAM) might not be possible, in shared memory systems, such as multi-core processors, the slowdown is mitigated by truly parallel access to private and shared caches, and thus the assumption is reasonable. We therefore follow this assumption in the same spirit.

In fact, for w equal to the regular word size (32 or 64 bits), the choice of w blocks of w bits each for the wide word ALU was judiciously made to provide the model with a feasible memory access implementation. w^2 lines to memory are well within the realm of the possible, as they are of the same order of magnitude (a factor of 2 or 8) as modern GPUs, some of which feature bus widths of 512 bits (e.g., FirePro W9100 [1] or Nvidia GeForce GTX 285 [21], see also [38, 39]). We note that a more general model could feature a wide word with k blocks of w bits each, where k is a parameter, which can be adjusted in reality according to the feasibility of implementation of parallel memory accesses. Although described for w blocks, the algorithms presented in this paper can easily be adapted to work with k blocks instead. Naturally, the speedups obtained would depend on the number of blocks assumed, but also on the memory bandwidth of the architecture. A practical implementation with a large number of blocks would likely suffer slowdowns due to congestion in the memory bus. We believe that an implementation with k equal to 32 or 64 can be realized with truly parallel memory access, leading to significant speedups.

Name	Input	Semantics
read_block	W, j, base	$W_j \leftarrow \text{MEM}[\text{base}+j]$
read_word	W, base	for all j in parallel: $W_j \leftarrow \text{MEM}[\text{base}+j]$
read_content	W, base	for all j in parallel: $W_j \leftarrow \text{MEM}[\text{base}+W_j]$
write_block	W, j, base	$\text{MEM}[\text{base}+j] \leftarrow W_j$
write_word	W, base	for all j in parallel: $\text{MEM}[\text{base}+j] \leftarrow W_j$
write_content	W, V, base	for all j in parallel: $\text{MEM}[\text{base}+V_j] \leftarrow W_j$

Table 1. Wide word memory access operations of the UW-RAM. MEM denotes regular RAM memory, which is indexed by addresses to words, and *base* is some base address.

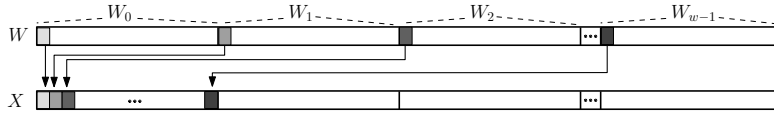


Fig. 2. The *compress* operation takes a wide word W whose set bits are restricted to the first bit of each block and compresses them to the first block of a wide word.

UW-RAM Subroutines We now describe some operations that will be used throughout the UW-RAM implementations that we describe in later sections. A procedure called *compress* serves to bring together bits from all blocks into one block in constant time, while a procedure called *spread* is the inverse function⁵. Both operations can be implemented by straightforward constant-depth circuits. We will also use parallel comparators, a standard technique used in word-RAM algorithms [24] (see details in Appendix A). Although these are all the subroutines that we need for the results in this paper, other operations of similar complexity could be defined if proved useful.

- **Compress:** Let W be a wide word in which all bits are zero except possibly for the first bit of each block. The compress operation copies the first bit of each block of W to the first block of a word X . I.e., if $X = \text{compress}(W)$, then $X[j] \leftarrow W_j[0]$ for $0 \leq j < w$, and $X[j] = 0$ for $j \geq w$ (see Fig. 2).
- **Spread:** This operation is the inverse of the compress operation. It takes a word W whose set bits are all in the first block and spreads them across blocks of a word X so that $X_j[0] \leftarrow W[j]$ for $0 \leq j < w$.

2.3 Relation to Other Models

There exist various models and architectures that exploit the execution of instructions on a large number of bits simultaneously. In Very Large Instruction Word (VLIW) architectures [17] several, possibly different instructions can be encoded in one wide word and executed in parallel. It is usually the compiler's

⁵ These operations are also known as PackSignBits and UnPackSignBits [37].

job to determine which instructions of a program can be executed safely in parallel. In contrast, in the UW-RAM model it is up to the algorithm designer to specify how parallelism in the ultra wide word should be used. In addition, the wide word can only execute one type of instruction at a time. In this sense, the UW-RAM is closer to a vector processor, in which a single instruction is executed on various data item, implementing SIMD parallelism. However, while vector processors operate on fields which are independent of each other, the ultra wide ALU in the UW-RAM is really one wide word of thousands of bits that treats its contents as one data object. An exception to this are the memory access instructions, which load and store data in blocks within the wide word so that the wide word ALU can interact with regular w -bit data. It is of course possible to use the ultra-wide word to implement a vectorized operation, however, as instructions in the UW-RAM operate on the entire word, it is up to the algorithm designer to deal with carries and other interference within fields. Moreover, the length of a field in the UW-RAM is variable, as it depends on the algorithm's choice. In that sense, the UW-RAM is a more flexible model.

Many modern processors support some form of SIMD parallelism with vectors of a small number of fields (e.g. Intel's SSE). Depending on the architecture, some of the available operations include inter-field instructions such as *shuffle* (which permutes fields in a vector), *pack* and *unpack* (equivalent to our compress and spread operations), inter-field shifts, or global sum (which sums all fields in the vector). The power of multimedia processors was studied by Thorup [37], who modeled these processors as vectors of k fields of ℓ bits each. Thorup showed that standard global operations on $(k \times \ell)$ -bit words can be implemented using vector instructions and inter-field operations in constant time, and argued that this enables the implementation of fundamental combinatorial algorithms such as sorting, hashing, and algorithms for minimum spanning trees on $(k \times \ell)$ -bit integers.

In contrast to Thorup's work, our main interest is in using the ultra wide word to deal with inputs of regular w -bit data objects and to speed up algorithms by being able to operate on more of these objects simultaneously. Moreover, we assume that the wide-word ALU supports the standard operations on the full word from the outset, with no need to simulate them using vector operations. Finally, we explore the consequences of indirect memory addressing at the field level, a feature that is not mentioned in Thorup's model.

The UW-RAM model can also be related to Multiple-Instruction-Multiple-Data (MIMD) models, and in particular to the PRAM. Although the UW-RAM ALU can only execute one instruction on the wide word, it is conceivable to devise a simulation of a PRAM algorithm on the UW-RAM. Each block of the wide word in the UW-RAM acts like a PRAM processor. Since the UW-RAM can only execute one type of instruction at a time, each parallel step of the PRAM algorithm is executed in $\lceil s/w \rceil$ steps on the UW-RAM, where s is the number of different instructions involved in the PRAM algorithm. For a constant number of different PRAM instructions and a non-constant number of UW-RAM blocks w , this simulation results in a constant overhead in time (compared to the PRAM

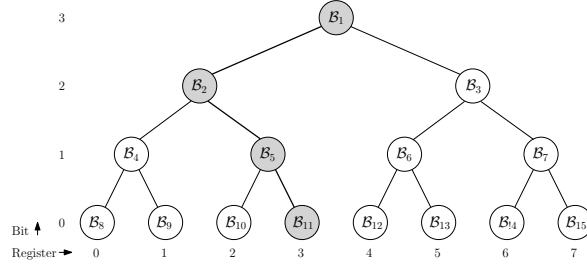


Fig. 3. Yggdrasil memory layout [10]: each node in a complete binary tree is an FS-RAM bit and registers are defined as paths from a leaf to the root. For example, register 3 contains bits \mathcal{B}_{11} , \mathcal{B}_5 , \mathcal{B}_2 , and \mathcal{B}_1 (shaded nodes).

algorithm running on $\Theta(w)$ processors). However, if such simulation were to be done in any practical implementation of these two models, the actual slowdown would be significant and most instructions would execute serially (as the number of different PRAM instructions is in the same order of magnitude as w). On the other hand, any UW-RAM algorithm that runs in time $t + q$, where q is the number of compress operations and t is the number of steps involved in the rest of the operations, can be simulated in time $O(t + q \log w)$ on a PRAM with w processors, as $\log w$ steps are necessary to simulate a compress operation.

Although simulations between the UW-RAM and other models exist, the idea of introducing the UW-RAM is to achieve larger speedups with word-RAM algorithms, keeping the programming techniques of this model. In practice, the implementations of PRAM algorithms are usually on asynchronous multi-cores, in which programmers must deal with concurrency issues. The advantage of our model is that we can avoid these issues while obtaining similar speedups to those of multi-cores.

3 Simulation of FS-RAM

In the standard RAM model of computation memory is organized in registers or words, each word containing a set of bits. Any bit in a word belongs to that word only. In contrast, in the FS-RAM model [18]—also known as Random Access Machine with Byte Overlap (RAMBO)—words can overlap, that is, a single bit of memory can belong to several words. The topology of the memory, i.e., a specification of which bits are contained in which words, defines a particular variant of the FS-RAM model. Variants of this model have been used to sidestep lower bounds for important data structure problems [10, 11].

We show how the UW-RAM can be used to implement memory access operations for any given FS-RAM of word size at most w bits in constant time. Thus, the time bounds of any algorithm in the FS-RAM model carry over directly to the UW-RAM. Note that each FS-RAM layout requires a different specialized hardware

implementation, whereas a UW-RAM architecture can simulate any FS-RAM layout without further changes to its memory architecture.

3.1 Implementing FS-RAM Operations in the UW-RAM

Let $\mathcal{B}_1, \dots, \mathcal{B}_B$ denote the bits of FS-RAM memory. A particular FS-RAM memory layout can be defined by the registers and the bits contained in them [9]. For example, in the *Yggdrasil* model in Fig. 3, $\text{reg}[0] = \mathcal{B}_8\mathcal{B}_4\mathcal{B}_2\mathcal{B}_1$, and in general $\text{reg}[i].\text{bit}[j] = \mathcal{B}_k$, where $k = \lfloor i/2^j \rfloor + 2^{m-j-1}$ ($m = 4$ in the example) [10].

In order to implement memory access operations on a given FS-RAM using the UW-RAM, we need to represent the memory layout of FS-RAM in standard RAM. Assume an FS-RAM memory of r registers of $b \leq w$ bits each and $B \leq br$ distinct FS-RAM bits. We assume that the FS-RAM layout is given as a table \mathcal{R} that stores, for each register and bit within the register, the number of the corresponding FS-RAM bit. Thus, if $\text{reg}[i].\text{bit}[j] = \mathcal{B}_k$, for some k , then $\mathcal{R}[i, j] = k$. We assume \mathcal{R} is stored in row major order. We simply store the value of each FS-RAM bit \mathcal{B}_i in a different w -bit entry of an array A in RAM, i.e., $A[i] = \mathcal{B}_i$. We could store more than one bit in each word of A ; however, this representation allows us to avoid having to serialize concurrent writes to the same word.

Given an index t of a register of an FS-RAM represented by \mathcal{R} , we can read the values of each bit of $\text{reg}[t]$ from RAM and return the b bits in a word. Doing this sequentially for each bit might take $O(b)$ time. Using the wide word we can take advantage of parallel reading and the compress operation to retrieve the contents of $\text{reg}[t]$ in constant time. Let $\text{reg}[t] = \mathcal{B}_{i_0} \dots \mathcal{B}_{i_{b-1}}$. The read operation first obtains the address in A of each bit of register t from \mathcal{R} . Then, it uses a content access to read the value of each bit \mathcal{B}_{i_j} into block W_j of W , thus assigning $W_j \leftarrow A[\mathcal{R}[t, j]]$. Finally, it applies one compress operation, after which the b bits are stored in W_0 . Algorithm 1 shows the read operation, which takes constant time. In order to implement the write operation $\text{reg}[t] \leftarrow \mathcal{B}_{i_0} \dots \mathcal{B}_{i_{b-1}}$ of FS-RAM, we first set $W_0 \leftarrow \mathcal{B}_{i_0} \dots \mathcal{B}_{i_{b-1}}$ and perform a spread operation to place each bit \mathcal{B}_j in block W_j . We then write the contents of each W_j in $A[\mathcal{R}[t, j]]$. Algorithm 2 shows this operation, which takes constant time as well.

Algorithm 1 fs-ram_read(t)

- 1: read_word($W, \mathcal{R}[t]$) $\{W_j \leftarrow \mathcal{R}[t, j]\}$
 - 2: read_content(W, A) $\{W_j \leftarrow A[\mathcal{R}[t, j]]\}$
 - 3: $W \leftarrow \text{compress}(W)$
 - 4: write_block($W, 0, \&ret$) $\{ret \leftarrow W_0\}$
 - 5: **return** ret
-

Since the read and write operations described above are sufficient to implement any operation that uses FS-RAM memory (any other operation is implemented in RAM), we have the following result.

Algorithm 2 $\text{fs_ram_write}(t, \mathcal{B} = \mathcal{B}_{i_0} \dots \mathcal{B}_{i_{b-1}})$

```
1: read_block( $W, 0, \mathcal{B}$ )  $\{W_0 \leftarrow \mathcal{B}\}$   
2:  $W \leftarrow \text{spread}(W)$   
3: read_word( $V, \mathcal{R}[t]$ )  $\{V_j \leftarrow \mathcal{R}[t, j]\}$   
4: write_content( $W, V, A$ )  $\{A[\mathcal{R}[t, j]] \leftarrow W_j\}$ 
```

Theorem 1. *Let \mathcal{R} be any FS-RAM memory layout of r registers of at most b bits each and B distinct FS-RAM bits, with $b \leq w$ and $\log B \leq w$. Let A be any FS-RAM algorithm that uses \mathcal{R} and runs in time T . Algorithm A can be implemented in the UW-RAM to run in time $O(T)$, using $rb + B$ additional words of RAM.*

Proof. Table \mathcal{R} indicating the FS-RAM bit identifier for each register and bit within register can be stored in rb words of RAM, while the values of each bit can be stored in B words of RAM. Since both fs_ram_read and fs_ram_write are constant time operations, any t -time operation that uses FS-RAM memory can be implemented in UW-RAM in the same time t . \square

3.2 Constant Time Priority Queue

Brodnik et al. [10] use the Yggdrasil FS-RAM memory layout to implement priority queue operations in constant time using $3M - 1$ bits of space ($2M$ of ordinary memory and $M - 1$ of FS-RAM memory), where M is the size of the universe. This problem has non-constant lower bounds for several models, including an $\Omega(\min\{\lg \lg M / \lg \lg \lg M, \sqrt{\lg N / \lg \lg N}\})$ lower bound in the RAM model when the memory is restricted to $N^{O(1)}$, where N is the number of elements in the set to be maintained [6]. For a universe of size $M = 2^m$, for some m , the Yggdrasil FS-RAM layout consists of $r = M/2$ registers of $b = \log M$ bits each, and $B = M - 1$ distinct FS-RAM bits (Fig. 3 is an example with $M = 16$). Thus, applying Theorem 1 we obtain the following result:

Corollary 1. *The discrete extended priority queue problem can be solved in the UW-RAM in $O(1)$ time per operation using $2M + w(M/2) \log M + w(M - 1)$ bits, thus in $O(M \log M)$ words of RAM.*

3.3 Constant Time Dynamic Prefix Sums

Brodnik et al. [11] use a modified version of the Yggdrasil FS-RAM to solve the dynamic prefix sums problem in constant time. This problem consists of maintaining an array A of size N over a universe of size M that supports the operations $\text{update}(j, d)$, which sets $A[j]$ to $A[j] \oplus d$, and $\text{retrieve}(j)$, which returns $\bigoplus_{i=0}^j A[i]$ [19, 11], where \oplus is any associative binary operation. This FS-RAM implementation sidesteps lower bounds on various models: there is an $\Omega(\log N)$ algebraic complexity lower bound [19] as well as under the semi-group model of computation [25], and an $\Omega(\log N / \log \log N)$ information-theoretic lower bound [19].

The result of Brodnik et al. [11] uses a complete binary tree on top of array A as leaves. The tree is similar to the one used in the priority queue problem, but it differs in that only internal nodes store any information and in that there are $m = \lceil \log M \rceil$ bits stored in each node. This tree is stored in a variant of the Yggdrasil memory called m -Yggdrasil, in which each register corresponds again to a path from a leaf to the root, but this time each node stores not only one bit but the m bits containing the sum of all values in the leaves of the left subtree of that node [11]. It is assumed that $nm \leq w$, where $n = \lceil \log N \rceil$ and w is the size of the word in bits. Thus, an entire path from leaf to root fits in a word and can be accessed in constant time. An update or retrieve operation consists of retrieving the values along a path in the tree and processing them in constant time using bit-parallelism and table lookup operations. The space used by the lookup table can be reduced at the expense of an increased time for the retrieve operation. In general, both operations can be supported in time $O(\iota + 1)$ with $(N - 1)m$ bits of m -Yggdrasil memory and $O(M^{n/2^\iota} \cdot m + m)$ bits of RAM, where ι is a trade-off parameter [11].

In order to represent the m -Yggdrasil memory in our model, we treat each bit of a node in the tree as a separate FS-RAM bit. Thus, the FS-RAM memory has $r = N$ registers of $b = nm$ bits each, and there are $B = (N - 1)m$ distinct bits to be stored. Hence, by Theorem 1 we have:

Corollary 2. *The operations update and retrieve of the dynamic prefix sums problem can be supported in the UW-RAM model in $O(\iota + 1)$ time with $O(M^{n/2^\iota} \cdot m + Nmnw)$ bits of RAM. For constant time operations ($\iota = 1$) the space is dominated by the first term, i.e., the space is $O(M^{\sqrt{\log N}})$ bits. For $\iota = \log \log N$, the time is $O(\log \log N)$ and the space is $O(Nmnw)$ bits.*

4 Dynamic Programming

In this section we describe UW-RAM implementations of dynamic programming algorithms for the subset sum, knapsack, and longest common subsequence problems. A word-RAM algorithm that only uses bit parallelism can be translated directly to the UW-RAM. The algorithm for subset sum is an example of this. In general, however, word-RAM algorithms that use lookup tables cannot be directly extended to w^2 bits, as this would require a mechanism to address $\Theta(w^2)$ -bit words in memory as well as lookup tables of prohibitively large size. Hence, extra work is required to simulate table lookup operations. The knapsack implementation that we present is a good example of such case.

4.1 Subset Sum

Given a set $S = \{a_1, a_2, \dots, a_n\}$ of nonnegative integers (weights) and an integer t (capacity), the subset sum problem is to find $S' \subseteq S$ such that $\sum_{a_i \in S'} a_i = t$. The optimization version asks for the solution of maximum weight which does not exceed t [14]. This problem is NP-hard, but it can be solved in pseudopolynomial

time via dynamic programming in $O(nt)$ time, using the following recurrence [7]: for each $0 \leq i \leq n$ and $0 \leq j \leq t$, $C_{i,j} = 1$ if and only if there is a subset of elements $\{a_1, \dots, a_i\}$ that adds up to j . Thus, $C_{0,0} = 1$, $C_{0,j} = 0$ for all $j > 0$, and $C_{i,j} = 1$ if $C_{i-1,j} = 1$ or $C_{i-1,j-a_i} = 1$ ($C_{i,j} = 0$ for any $j < 0$). The problem admits a solution if $C_{n,t} = 1$.

Pisinger [35] gives an algorithm that implements this recursion in the word-RAM with word size w by representing up to w entries of a row of C . Using bit parallelism, w bits of a row can be updated simultaneously in constant time from the entries of the previous row: C_i is updated by computing $C_i = (C_{i-1} \mid (C_{i-1} \gg a_i))$ (which might require shifting words containing C_{i-1} first by $\lfloor a_i/w \rfloor$ words and then by $a_i - \lfloor a_i/w \rfloor$) [35]. Assuming $w = \Theta(\log t)$, this approach leads to an $O(nt/\log t)$ time solution in $O(t/\log t)$ space. The actual elements in S' that form the solution can be recovered with the same space and time bounds with a recursive technique by Pferschy [34].

This algorithm can be implemented directly in the UW-RAM: entries of row C_i are stored contiguously in memory; thus, we can load and operate on w^2 bits in $O(1)$ time when updating each row. Hence, the UW-RAM implementation runs in $O(nt/\log^2 t)$ time using the same $O(t/\log t)$ space (number of w -bit words).

4.2 Knapsack

Given a set S of n elements with weights and values, the knapsack problem asks for a subset of S of maximum value such that the total weight is below a given capacity bound b . Let $S = \{(w_i, v_i)\}_{i=1}^n$, where w_i and v_i are the weight and value of the i -th element. Like subset sum, this problem is NP-hard but can be solved in pseudopolynomial time using the following recurrence [7]: let $C_{i,j}$ be the maximum value of a solution containing elements in the subset $S_i = \{(w_k, v_k)\}_{k=1}^i$ with maximum capacity j . Then, $C_{0,j} = 0$ for all $0 \leq j \leq b$, and $C_{i,j} = \max\{C_{i-1,j}, C_{i-1,j-w_i} + v_i\}$. The value of the optimal solution is $C_{n,b}$. This leads to a dynamic program that runs in $O(nb)$ time.

The word-RAM algorithm by Pisinger [35] represents partial solutions of the dynamic programming table with two binary tables g and h and operates on $O(w)$ entries at a time. More specifically, $g_{i,u} = 1$ and $h_{i,v} = 1$ if and only if there is a solution with weight u and value v that is not dominated by another solution in $C_{i,*}$ (i.e., there is no entry $C_{i,u'}$ such that $u' < u$ and $C_{i,u'} \geq v$). Pisinger shows how to update each entry of g and h with a constant time procedure, which can be encoded as a constant size lookup table T . A new lookup table T^α is obtained as the product of α times the original table T . Thus, α entries of g and h can be computed in constant time. Setting $\alpha = w/10$, an entire row of g and h can be computed in $O(m/w)$ time and $O(m/w)$ space [35], where m is the maximum of the capacity b and the value of the optimal solution⁶. The optimal solution can then be computed in $O(nm/w)$ time.

⁶ This value is not known in advance, though an upper bound of at most twice the optimal value can be used [35, 16].

Compared to the subset sum algorithm, which relies mainly on bit-parallel operations, this word-RAM algorithm for knapsack relies on precomputation and use of lookup tables to achieve a w -fold speedup. While we cannot precompute a composition of $\Theta(w^2)$ lookup tables to compute $\Theta(w^2)$ entries of g and h at a time, we can use the same tables with $\alpha = w/10$ as in Pisinger's algorithm and use the *read_content* operation of the UW-RAM to make w simultaneous lookups to the table. Since the entries in a row i of h and g depend only on entries in row $i - 1$, then there are no dependencies between entries in the same row.

One difficulty is that in order to compute the entries in row i in parallel we must first preprocess row $i - 1$ in both h and g , such that we can return the number of one bits in both $g_{i-1,0}, \dots, g_{i-1,j}$ and $h_{i-1,0}, \dots, h_{i-1,j}$ in $O(1)$ time for any column $j \in \{0, m - 1\}$. That is, the prefix sums of the one bits in row $i - 1$. Note that this is *not* the same as the dynamic problem described in Section 3.3, but it is a static prefix sums problem. Furthermore, since the algorithm is the same for both g and h , we describe the computation for g alone.

Static Prefix Sums We divide g_{i-1} in blocks of w contiguous bits and compute the number of ones in each block $g_{i-1,k}, \dots, g_{i-1,k+w-1}$ for $k \in \{0, w, 2w, \dots, \lfloor m/w \rfloor w\}$ using a lookup table. We store the results in an array \mathcal{A} of length $\lceil m/w \rceil$, with $\mathcal{A}[k]$ storing the number of ones in the k -th block. Next, we compute the prefix sums \mathcal{A}' of \mathcal{A} in two steps. We divide \mathcal{A} in subarrays of w consecutive entries. Let \mathcal{A}_i denote the subarray $\mathcal{A}[iw, iw + w - 1]$, for $i \in \{0, 1, \dots, \lceil |\mathcal{A}|/w \rceil - 1\}$.

The first step is to compute the prefix sums \mathcal{A}'_i of each subarray \mathcal{A}_i , i.e. $\mathcal{A}'_i[k] = \sum_{j=0}^k \mathcal{A}_i[j]$. Using the w blocks of a wide word, we can operate on w entries at a time. Consider the first w consecutive subarrays $\mathcal{A}_0, \mathcal{A}_1, \dots, \mathcal{A}_{w-1}$. In order to compute $\mathcal{A}'_0, \dots, \mathcal{A}'_{w-1}$, for each $0 \leq k \leq w - 1$, we use the i -th block of the wide work to compute $\mathcal{A}'_i[k]$, thus computing the entries for all $0 \leq i \leq w - 1$ simultaneously. Each entry is computed in constant time, since

$$\mathcal{A}'_i[k] = \begin{cases} \mathcal{A}'_i[k-1] + \mathcal{A}_i[k] & \text{if } k > 0, \\ \mathcal{A}_i[k] & \text{otherwise.} \end{cases}$$

Hence, we can compute the prefix sums of w subarrays in $O(w)$ time. After computing the first w subarrays we continue with the second group, and so on. Thus, we compute all prefix sums of the $O(|\mathcal{A}|/w)$ subarrays in $O(|\mathcal{A}|/w)$ time.

The second step is to update each subarray of \mathcal{A}' by adding to each entry the last entry of the previous subarray. I.e., we set $\mathcal{A}'_i[k] = \mathcal{A}'_i[k] + \mathcal{A}'_{i-1}[w-1]$ for all $i = 1, \dots, \lceil |\mathcal{A}'|/w \rceil - 1$ (in increasing value of i). This can also be done for w entries at once, but this time we use the blocks of the wide word to update all entries of one subarray simultaneously. Thus, sequentially for each $i = 1, \dots, \lceil |\mathcal{A}'|/w \rceil - 1$ we update \mathcal{A}'_i in $O(1)$ time, and hence \mathcal{A}' is updated in $O(|\mathcal{A}|/w)$ time.

At this point, \mathcal{A}' contains the prefix sums of \mathcal{A} , and took $O(|\mathcal{A}|/w) = O(m/w^2)$ time to compute. Fig. 4 shows an example of this procedure.

g	=	100		011		110		111		001		101		100		11
\mathcal{A}	=	1		2		2		3		1		2		1		2
		\mathcal{A}_0						\mathcal{A}_1						\mathcal{A}_2		
Step 1																
$\mathcal{A}'(1)$	=	<u>1</u>						<u>3</u>						<u>1</u>		
$\mathcal{A}'(2)$	=	1		<u>3</u>				3		<u>4</u>				1		<u>3</u>
$\mathcal{A}'(3)$	=	1		3		<u>5</u>		3		4		<u>6</u>		1		3
Step 2																
$\mathcal{A}'(1)$	=	1		3		5		<u>8</u>		<u>9</u>		<u>11</u>		1		3
$\mathcal{A}'(2)$	=	1		3		5		8		9		11		<u>12</u>		<u>14</u>

Fig. 4. Example of computing prefix sums in the UW-RAM with $w = 3$ and $m = 23$. Numbers in parenthesis indicate the parallel step number when computing \mathcal{A}' and underlined entries indicate the entries computed in that step.

Let f be the number of ones in $g_{i-1, \lfloor j/w \rfloor}, \dots, g_{i-1, j}$, which can be computed using the lookup table. To compute the number of ones in $g_{i-1, 0}, \dots, g_{i-1, j}$ we return $f + \mathcal{A}'[\lfloor j/w \rfloor]$.

Then, each row of g and h takes $O(m/w^2)$ time to compute, and since there are n rows, the total time to compute g and h (and hence the optimal solution) on the UW-RAM is $O(nm/w^2)$. This achieves a w -fold speedup over Pisinger's word-RAM solution.

4.3 Generalizations of Subset Sum and Knapsack Problems

Pisinger [35] uses the techniques of the word-RAM algorithm for subset sum and knapsack to obtain a word-RAM algorithm for computing a path in a layered network: given a graph $G = (V, E)$, a source $s \in V$ and a terminal $t \in V$, and a weight for each edge, is there a path of weight b from s to t ? Again, this algorithm translates directly to a UW-RAM algorithm, thus yielding a w -fold speedup over the word-RAM algorithm. Pisinger further uses the algorithms for the problems above to implement word-RAM solutions for other generalizations of subset sum and knapsack problems, such as: the bounded subset sum and knapsack problems (each element can be chosen a bounded number of times), the multiple choice subset sum and knapsack problems (the set of numbers is divided in classes and the target sum must be matched with one number of each class), the unbounded subset sum and knapsack problems (each element can be chosen an arbitrary number of times), the change-making problem, and, finally, the two-partition problem. UW-RAM implementations for all these generalizations are direct and yield a w -fold speedup over the word-RAM algorithms (recall that $w = \Omega(\log n)$).

4.4 Longest Common Subsequence

The final dynamic programming problem we examine is that of computing the longest common subsequence (LCS) of two string sequences (Definition 1).

Definition 1. [LCS] Given a sequence of symbols $X = x_1x_2 \dots x_m$, a sequence $Z = z_1z_2 \dots z_k$ is a subsequence of X if there exists an increasing sequence of indices i_1, i_2, \dots, i_k such that for all $1 \leq j \leq k$, $x_{i_j} = z_j$ [14]. Let Σ be a finite alphabet of symbols, and let $\sigma = |\Sigma|$. Given two sequences $X = x_1x_2 \dots x_m$ and $Y = y_1y_2 \dots y_n$, where $x_i, y_j \in \Sigma$, the Longest Common Subsequence problem asks for a sequence $Z = z_1z_2 \dots z_k$ of maximum length such that Z is a subsequence of both X and Y .

This problem can be solved via a classic dynamic programming algorithm in $O(nm)$ time [14]. We describe a UW-RAM algorithm for LCS based on an algorithm by Masek and Paterson [31]. We note that there exist other approaches to solving the LCS problem with bit-parallelism (e.g., [15]) that could also be adapted to work in the UW-RAM. The approach we show here is a good example of bit parallelism combined with the parallel lookup power of the model, which we use to implement the Four Russians technique.

The base algorithm, which mainly relies on bit parallelism, leads to Theorem 2. We then extend the algorithm with the Four Russians technique to achieve further speedups, obtaining Theorem 3.

Theorem 2. The length of the LCS of two strings X and Y over an alphabet of size σ , with $|X| = m$ and $|Y| = n$, can be computed in the UW-RAM in $O(\frac{nm}{w^2} \log \sigma + m + n)$ time and $O(\frac{\min(n,m)}{w} \log \sigma)$ words in addition to the input.

Theorem 3. The length of the LCS of two strings X and Y of length n over an alphabet of size σ can be computed in the UW-RAM in $O(n^2 \log^2(\sigma)/w^3 + n \log(\sigma)/w)$ time. For $\sigma = O(1)$ and $w = \Theta(\log n)$ this time is $O(n^2/\log^3 n)$.

Let $c_{i,j}$ denote the length of the LCS of $X[1..i] = x_1x_2 \dots x_i$ and $Y[1..j] = y_1y_2 \dots y_j$. Then the following recurrence allows us to compute the length of the LCS of X and Y [14]:

$$c_{i,j} = \begin{cases} 0, & \text{if } i = 0 \text{ or } j = 0 \\ c_{i-1,j-1} + 1, & \text{if } x_i = y_j \\ \max\{c_{i,j-1}, c_{i-1,j}\}, & \text{otherwise.} \end{cases} \quad (1)$$

The length of the LCS is $c_{m,n}$, which can be computed in $O(mn)$ time. Consider an $(m+1) \times (n+1)$ table C storing the values $c_{i,j}$. The idea of the UW-RAM algorithm is to compute various entries of this table in parallel. We assume $w = \Theta(\max\{\log n, \log m\})$.

Let d_k denote the values in the k -th diagonal of table C , this is $d_k = \{c_{i,j} | i + j = k\}$. Since a value in a cell $i, j > 0$ depends only on the values of cells $(i-1, j)$, $(i-1, j-1)$ and $(i, j-1)$, all values in the same diagonal d_k can be computed in parallel. Thus, we use the wide word to compute various entries of a diagonal in constant time. Since each value in the cell might use up to $\min\{\log n, \log m\}$ bits, each value might use up to an entire block of the wide word (if $\log m = \Theta(\log n)$); thus, w cells can be computed in parallel. Since the total number of cells is $O(mn)$ and the critical path of the table has $m + n + 1$

LCS	j	1	2	3	4	5	6
		a	a	b	b	b	a
i		0	0	0	0	0	0
1 a		0	1	1	1	1	1
2 b		0	1	1	2	2	2
3 b		0	1	1	2	3	3
4 a		0	1	2	2	3	4
5 b		0	1	2	3	3	4

H	j	1	2	3	4	5	6
		a	a	b	b	b	a
i		0	0	0	0	0	0
1 a		1	0	0	0	0	0
2 b		1	0	1	0	0	0
3 b		1	0	1	1	0	0
4 a		1	1	0	1	0	1
5 b		1	1	0	1	0	1

V	j	1	2	3	4	5	6
		a	a	b	b	b	a
i		0	1	1	1	1	1
1 a		0	1	1	1	1	1
2 b		0	0	0	1	1	1
3 b		0	0	0	0	1	1
4 a		0	0	1	0	0	1
5 b		0	0	0	1	0	1

Fig. 5. Dynamic programming tables for the LCS and horizontal and vertical differences for $X = abbab$ and $Y = aabbba$.

cells, this approach takes $O(mn/w + m + n)$ parallel time, resulting in a speedup of w . However, we can obtain better speedups by using fewer bits per entry of the table, which enables us to operate on more values in parallel. For this sake, instead of storing the actual values of the partial longest common subsequences, we store differences between consecutive values as described in [31] for the related string edit distance problem.

Let V and H denote the tables of vertical and horizontal differences of values in C , respectively. Entries in these tables are defined as $V_{i,j} = c_{i,j} - c_{i-1,j}$ and $H_{i,j} = c_{i,j} - c_{i,j-1}$ for $1 \leq i \leq m$ and $1 \leq j \leq n$. Fig. 5 shows the tables C , V , and H for an example pair of input sequences. We adapt Corollary 1 in [31] for the computation of V and H :

Proposition 1. *Let $[x_i = y_j] = 1$ if $x_i = y_j$ and 0 otherwise. Then, $V_{i,j} = \max\{[x_i = y_j] - H_{i-1,j}, 0, V_{i,j-1} - H_{i-1,j}\}$ and $H_{i,j} = \max\{[x_i = y_j] - V_{i,j-1}, 0, H_{i-1,j} - V_{i,j-1}\}$.*

Proof. Directly from Recurrence (1) we obtain $V_{i,j} = 1 - H_{i-1,j}$ if $x_i = y_j$ and $V_{i,j} = \max\{0, V_{i,j-1} - H_{i-1,j}\}$ otherwise. Similarly, $H_{i,j} = 1 - V_{i,j-1}$ if $x_i = y_j$ and $H_{i,j} = \max\{0, H_{i-1,j} - V_{i,j-1}\}$ otherwise. It is easy to verify from the definition of longest common subsequence and Recurrence (1) that $0 \leq H_{i,j} \leq 1$ and $0 \leq V_{i,j} \leq 1$ for all i, j , which implies that the maximum in $\max\{[x_i = y_j] - H_{i-1,j}, 0, V_{i,j-1} - H_{i-1,j}\}$ and $\max\{[x_i = y_j] - V_{i,j-1}, 0, H_{i-1,j} - V_{i,j-1}\}$ is equal to the first term if $x_i = y_j$ and to the second or third terms otherwise. \square

We compute tables H and V according to Proposition 1 diagonal by diagonal using bit parallelism in the wide word. Assume an alphabet $\Sigma = \{0, 1, 2, \dots, \sigma - 1\}$ with $\lceil \log \sigma \rceil \leq w - 1$. Although all entries in tables H and V are either 0 or 1, we will use fields of $O(\log \sigma)$ bits to store these values, since we can only compare at most $w^2 / \log \sigma$ symbols simultaneously in the wide word. We divide the wide word W in f -bit fields with $f = \max(\lceil \log \sigma \rceil, 2) + 1$. Each field will be used to store both symbols and intermediate results for the computation of the diagonals of H and V , plus an additional bit to serve as a test bit in order to implement fieldwise comparisons as described in Appendix A. We require at least 3 bits because although all entries in tables H and V use one bit, intermediate results

in calculations can result in values of -1. Thus, we require 2 bits to represent values -1, 0, and 1, and a test or sentinel bit to prevent carry bits resulting from subtractions to interfere with neighboring fields. We represent -1 in two's complement. It is not hard to extend the techniques for comparisons and maxima to the case of positive and negative numbers [24].

Let H_k and V_k denote the k -th diagonal of H and V , respectively, i.e., $H_k = \{H_{i,j} | i+j = k\}$ and $V_k = \{V_{i,j} | i+j = k\}$. Consider table H . We will operate with each diagonal H_k using $\lceil |H_k|/\ell \rceil$ wide words, where $\ell = \lfloor w^2/f \rfloor$. Let $f_0, \dots, f_{\ell-1}$ denote the fields within a wide word in increasing order of bit significance. In each wide word, cells of H_k will be stored in increasing order of column, i.e., if $H_{i,j}$ is stored in field f_r , then f_{r+1} stores $H_{i-1,j+1}$. In order to compute each diagonal we must compare the relevant entries of strings X and Y . We assume that each symbol of X and Y is stored using $\lceil \log \sigma \rceil + 1$ bits (including the test bit) and that X is stored in reverse order. X and Y can be preprocessed in $O(m+n)$ to arrange this representation, which will allow us to do constant-time parallel comparisons of symbols for each diagonal loading contiguous words of memory in wide words.

Consider a diagonal H_k . Assume that the entire diagonal fits in a word W . This will not be the case for most diagonals, but we describe the former case for simplicity. The latter case is implemented as a sequence of steps updating portions of the diagonal that fit in a wide word. We update the entries of H_k as follows:

1. We load the symbols of the relevant substrings of X and Y into words W_X and W_Y , with the substring of X in reverse order. More specifically, for a diagonal k , $W_Y = y_{j_1}y_{j_1+1} \dots y_{j_2}$, where $j_1 = k - \min(|X|, k-1)$ and $j_2 = \min(|Y|, k)$, and $W_X = x_{i_2}x_{i_2-1} \dots x_{i_1}$ with $i_2 = k - j_1$ and $i_1 = k - j_2$. We subtract W_Y from W_X , mask out all non-zero results and write a 1 in each field that resulted in 0. We store the resulting word in W_{eq} , where each field corresponding to a cell (i, j) stores a 1 if $x_i = y_j$ and a 0 otherwise (this can be implemented through comparisons as described in Appendix A).
2. We load V_{k-1} into a word W_V and subtract it from W_{eq} to obtain $[a_i = b_j] - V_{i,j-1}$ for all i, j in H_k simultaneously and store the result in W_1 .
3. We load H_{k-1} into a word W_H and subtract W_V from it to obtain $H_{i-1,j} - V_{i,j-1}$ for all i, j in H_k , storing the result in W_2 .
4. Finally, using fieldwise comparisons, we obtain the fieldwise maximum of W_1, W_2 and the word $\mathbf{0}$. The resulting word is H_k .

All the operations described above can be implemented in constant time. The procedure to compute V_k is analogous. Note that the entries corresponding to base cases in the first row and column in the LCS table correspond to the base cases of the horizontal and vertical vectors, respectively. When computing diagonals H_k with $k \leq n+1$ and V_k with $k \leq m+1$, the entries corresponding to base cases are not computed from previous diagonals but should be added appropriately at the end of H_k and beginning of V_k . Example 1 shows how to compute H_6 from H_5 and V_5 (in gray) in Fig. 5 with the above procedure.

Example 1. Let $X = abbab$ and $Y = aabbba$ be two strings. Fig. 5 shows the entries of the dynamic programming table for computing the LCS of X and Y , as well as the values of horizontal and vertical differences.

In this example $\sigma = 2$, thus we use one bit for each symbol ('a'=0, 'b'=1), but we use $f = 3$ bits per field. Consider the diagonal H_6 in table H (in dark gray). We now illustrate how to obtain H_6 from H_5 and V_5 (in light gray). In what follows we represent the number in each field in decimal and do not include the details of fieldwise comparison and maxima.

W_X	= 1 0 1 1 0	($=x_5x_4x_3x_2x_1$)
W_Y	= 0 0 1 1 1	($=y_1y_2y_3y_4y_5$)
W_{eq}	= 0 1 1 1 0	($W_{eq}[f(j-1)] = 1 \Leftrightarrow x_{ H_5 -j} = y_j$)
V_5	= 0 0 0 1 1	
$W_1 = W_{eq} - V_5$	= 0 0 1 0 -1	
H_5	= 1 0 1 0 0	
$W_2 = H_5 - V_5$	= 1 0 1 -1 -1	
$\max\{W_1, W_2, \mathbf{0}\}$	= 1 1 1 0 0	
H_6	= 1 1 1 0 0	(last 0 is the base case)

Once all diagonals are computed, the final length of the longest common subsequence of X and Y can be simply computed by (sequentially) adding the values of the last row of H or the values of last column of V (which can be done while computing H and V). The entire procedure is described in Algorithm 3 and leads to Theorem 2:

Theorem 2. *The length of the LCS of two strings X and Y over an alphabet of size σ , with $|X| = m$ and $|Y| = n$, can be computed in the UW-RAM in $O(\frac{nm}{w^2} \log \sigma + m + n)$ time and $O(\frac{\min(n,m)}{w} \log \sigma)$ words in addition to the input.*

Proof. A diagonal of H and V of length ℓ entries can be computed in time $O(\ell \log \sigma / w^2 + 1)$. Adding this time over all $m + n$ diagonals yields the total time. For the space, each diagonal is represented in $\lceil \ell f / w^2 \rceil$ wide words, where $f = O(\log \sigma)$ is the number of bits per field. Since we can compute each diagonal H_k and V_k using only H_{k-1} and V_{k-1} , we only need to store 4 diagonals at any given time. Since the maximum length of a diagonal is $\min(n, m) + 1$ and each wide word can be stored in w regular words of memory, the result follows. \square

Recovering a Longest Common Subsequence It is known that given a dynamic programming table storing the values of the LCS between strings X and Y , one can recover the actual subsequence by starting from $c_{m,n}$ and following the path through the cells corresponding to the values used when computing each value $c_{i,j}$ according to Recurrence (1): if $x_i = y_j$, then we add x_i to the LCS and continue with cell $(i-1, j-1)$; otherwise the path follows the cell corresponding to the maximum of $c_{i-1,j}$ or $c_{i,j-1}$. Although Algorithm 3 does not compute the actual LCS table, a path of an LCS can be easily computed using tables H and V . The path starts at cell (m, n) (of either table). Then, to

Algorithm 3 LCS-length($X, Y, m = |X|, n = |Y|, \sigma$)

```
1:  $f \leftarrow \max(\lceil \log \sigma \rceil, 2) + 1$  {field length in bits}
2:  $H_1^1 \leftarrow \mathbf{0}$   $\{H_{0,1} = 0\}$ 
3:  $V_1^1 \leftarrow \mathbf{0}$   $\{V_{1,0} = 0\}$ 
4:  $\text{length} \leftarrow 0$  {length of longest common subsequence}
5: for  $k = 2$  to  $m + n$  do
6:    $\ell \leftarrow \min(n, k - 1) + \min(m, k - 1) - k + 1$  {length of diagonal}
7:    $j_1 \leftarrow k - \min(m, k - 1)$  {indices of relevant substrings of  $X$  and  $Y$ }
8:    $j_2 \leftarrow \min(n, k)$ 
9:    $i_2 \leftarrow k - j_1$ 
10:   $i_1 \leftarrow k - j_2$ 
11:   $j \leftarrow j_1$ 
12:   $i \leftarrow i_2$ 
13:   $s \leftarrow \lceil \ell f / w^2 \rceil$  {number of wide words per diagonal}
14:  for  $t = 1$  to  $s$  do
15:     $j' \leftarrow \min(j + s - 1, j_2)$ 
16:     $i' \leftarrow \max(i + s - 1, i_1)$ 
17:     $W_Y \leftarrow Y[j..j']$ 
18:     $W_X \leftarrow X[i..i']$  {substring of  $X$  is in reverse order}
19:     $W_{eq} \leftarrow \text{equal}(W_X, W_Y)$ 
20:     $W_1 \leftarrow W_{eq} - V_{k-1}^t$ 
21:     $W_2 \leftarrow H_{k-1}^t - V_{k-1}^t$ 
22:     $H_k^t \leftarrow \max(W_1, W_2, \mathbf{0})$  {base case is implicitly added at rightmost field}
23:     $W_1 \leftarrow W_{eq} - H_{k-1}^t$ 
24:     $W_2 \leftarrow V_{k-1}^t - H_{k-1}^t$ 
25:     $V_k^t \leftarrow \max(W_1, W_2, \mathbf{0})$ 
26:    if  $t = 1$  AND  $k \leq m + 1$  then
27:       $V_k^t \leftarrow V_k^t >> f$  {add 0 in the first field for the base case}
28:     $i \leftarrow i' + 1$ 
29:     $j \leftarrow j' + 1$ 
30:    if  $t = 1$  AND  $k \geq m + 1$  then
31:       $\text{length} \leftarrow \text{length} + H_k^1[0..f - 1]$  {length = length +  $H_{m,k-m}$ }
32: return length
```

continue from a cell (i, j) , if $x_i = y_j$, then x_i is part of the LCS, and we continue with cell $(i - 1, j - 1)$; otherwise, if $H_{i,j} = 1$ and $V_{i,j} = 0$, then we continue with cell $(i - 1, j)$, and if $H_{i,j} = 0$ and $V_{i,j} = 1$, we continue with cell $(i, j - 1)$ (and with any of the two if $H_{i,j} = V_{i,j} = 0$). This can be easily done in $O(m + n)$ time if all diagonals of tables V and H are kept in memory while computing the LCS length in Algorithm 3. This would require Algorithm 3 to use $O(nmw / \log \sigma)$ words of memory to store all diagonals.

Four Russians Technique The computation of the longest common subsequence in the UW-RAM can be made even faster by combining the diagonal-by-diagonal order of computation described above with the Four Russians technique. The Four Russians technique [3] was used by Masek and Paterson to speedup

the computation of the string edit problem (and also the LCS) in a RAM with indirect addressing [31]. The technique consists of dividing the dynamic programming table in blocks of size $t \times t$ cells. In a precomputation phase, all possible blocks are computed and stored as a data structure indexed by the first row and column of each block. The LCS can be then computed by looking up relevant values of the table one block at a time using the data structure. In a RAM with indirect addressing and under a suitable value of t , the last row and column of a block can be obtained by looking up the entry corresponding to the first row and column of that block in constant time. This technique yields a speedup of $O(t^2)$ with respect to computing all cells in the table, for a total time of $O(n^2/t^2)$ (for two strings of length n) plus the time for the precomputation of all blocks. By setting $t = O(\log n)$ (for a constant alphabet size) and encoding the table with difference vectors, the precomputation time can be absorbed by the time to compute the main table (see [31, 23] for a more detailed description of the technique).

We can use the power of parallel memory accesses of the UW-RAM to speedup the computation of the LCS even further by looking up blocks in parallel, in a similar fashion to the diagonal-by-diagonal approach described above. For simplicity, assume $m = n$. Using the same encoding for H and V , we first precompute all possible blocks of H and V of size $t \times t$. Since a block is completely determined by its first column and row, whose values are in $\{0, 1\}$, and the two substrings of length t (over an alphabet of size σ), there are $O((2\sigma)^{2t})$ possible blocks. Note that we can encode each cell now with one bit, since we do not need to do symbol comparisons in parallel. Each block can be computed in $O(t^2)$ time with the standard sequential algorithm, so the precomputation time is $O((2\sigma)^{2t}t^2)$. We set $t = \log_{2\sigma} n/2$, and thus the precomputation time is $O(n \log^2 n)$ [23]. Since $t \leq w/2$, we can use each block of the wide word to lookup the entry for each block by using a parallel lookup operation. Thus, as described previously, we can compute tables H and V in diagonals of blocks, computing $\min(\ell, w)$ blocks simultaneously in a diagonal of length ℓ blocks. There are $(n/t)^2$ blocks to compute and the critical path of the table has length n/t blocks. Therefore, the computation of H and V can be carried out in time $O(n^2/(t^2w) + n/t) = O(n^2 \log^2 \sigma / w^3 + n \log \sigma / w)$, since $t = \Theta(w / \log \sigma)$. This result is summarized by Theorem 3:

Theorem 3. *The length of the LCS of two strings X and Y of length n over an alphabet of size σ can be computed in the UW-RAM in $O(n^2 \log^2(\sigma)/w^3 + n \log(\sigma)/w)$ time. For $\sigma = O(1)$ and $w = \Theta(\log n)$ this time is $O(n^2/\log^3 n)$.*

5 String Searching

Another example of a problem where a large class of algorithms can be sped up in the UW-RAM is string searching. Given a text T of length n and a pattern P of length m , both over an alphabet Σ , string searching consists of reporting all the occurrences of P in T . We focus here on on-line searching, this is, with

no preprocessing of the text (though preprocessing of the pattern is allowed), and we assume in general that $n \gg m$. We use two classic algorithms for this problem to illustrate different ways of obtaining speedups via parallel operations in the wide word. More specifically, we obtain speedups of $w = \Omega(\log n)$ for UW-RAM implementations of the Shift-And and Shift-Or algorithms [4, 40], and the Boyer-Moore-Horspool algorithm [28]. For a string S , let $S[i]$ denote its i -th character, and let $S[i..j]$ be the substring of S from position i to j . Indices start at 1.

5.1 Shift-And and Shift-Or

The Shift-And and Shift-Or algorithms keep a sliding window of length m over the text T . On a window at substring $T[i - m + 1..i]$, the algorithms keep track of all prefixes of P that match a suffix of $T[i - m + 1..i]$. Thus, if at any time there is one such prefix of length $|P|$, then an occurrence is reported at $T[i - m + 1]$. This is equivalent to running the $(m + 1)$ -state non-deterministic automaton that recognizes P starting from every position of T . For a window $T[i - m + 1..i]$ in T , the j -th state of the automaton ($0 \leq j \leq m$) is active if and only if $P[1..j] = T[i - j + 1..i]$. These algorithms represent the automaton as a bit vector and update the active states using bit-parallelism. Their running time is $O(mn/w + n)$, achieving linear time on the size of the text for small patterns. More specifically, the Shift-And algorithm keeps a bit vector $\mathbf{v} = b_1b_2 \dots b_m$, where $b_j = 1$ whenever the j -th state is active. If \mathbf{v}_i represents the automaton for the window ending at $T[i]$, then $\mathbf{v}_{i+1} = ((\mathbf{v}_i \gg 1) \mid 1) \& Y[T[i + 1]]$, where $Y[\sigma]$ is a bit vector with set bits in the positions of the occurrences of σ in P . The OR with a 1 corresponds to the initial state always being active to allow a match to start at any position. The Shift-Or algorithm is similar but it saves this operation by representing active states with zeros instead of ones.

We describe in two UW-RAM algorithms for Shift-And that illustrate different techniques, noting that the UW-RAM implementation of Shift-Or is analogous. We obtain the following theorem:

Theorem 4. *Given a text T of length n and a pattern P of length m , we can find the occ occurrences of P in T in the UW-RAM in time $O(nm/w^2 + n/w + \text{occ})$.*

w^2 -bit Automaton The straightforward way of taking advantage of the wide word when implementing Shift-And is to use the entire wide word for bit vectors. We first compute the mask array $Y[\sigma]$ for each $\sigma \in \Sigma$ and store each w^2 -bit vector in contiguous words of memory starting at address $Y + \sigma$. Then the code of the UW-RAM is essentially the same as the original code, replacing all references to the array Y with memory access operations for the wide word: assuming $m \leq w^2$, reading from and writing to $Y[\sigma]$ implemented by `read_word($W, Y + \sigma$)` and `write_word($W, Y + \sigma$)`, for some word W . Otherwise, bit vectors are represented in $\lceil m/w^2 \rceil$ wide words (and stored in memory in $\lceil m/w^2 \rceil w$ words). The rest of the operations are done on registers, and constants are part of the

Algorithm 4 Shift-And($T, P, n = |T|, m = |P|, \Sigma$)

```
1: {Preprocessing}
2: for each  $\sigma \in \Sigma$  do
3:    $Y[\sigma] \leftarrow \mathbf{0}$ 
4: for  $j = 1$  to  $m$  do
5:    $Y[P[j]] \leftarrow Y[P[j]] \mid (1 \gg (j - 1))$ 
6: {Search}
7:  $V \leftarrow \mathbf{0}$ 
8:  $C \leftarrow 1 \gg (m - 1)$ 
9: for  $i = 1$  to  $n$  do
10:   $V = ((V \gg 1) \mid 1) \& Y[T[i]]$ 
11:  if  $V \& C \neq 0$  then
12:    report an occurrence at  $i - m + 1$ 
```

precomputation. The pseudocode for this algorithm is shown in Algorithm 4, which assumes $m \leq w^2$ and is based on the pseudocode for Shift-And given in [33, Chapter 2.2.2]. Since we can now update v in $O(m/w^2 + 1)$ time, the running time of Algorithm 4 is $O(nm/w^2 + n)$. Thus, compared to the original algorithm, the UW-RAM algorithm achieves a speedup of w when $m \geq w^2$, and a speedup of $\lceil m/w \rceil$ otherwise (no speedup is achieved for $m \leq w$).

Lemma 1. *When implemented in the UW-RAM, the Shift-And and Shift-Or algorithms for searching a pattern of length m in a text of length n have a running time of $O(nm/w^2 + n)$, achieving a w -fold speedup over word-RAM implementations when $m \geq w^2$.*

w -bit Parallel Automata Another way of using the wide word to speedup the Shift-And algorithm is to take advantage of the parallel memory access operations of the UW-RAM to perform w parallel searches on disjoint portions of the text. This is done by using each block of a wide word to represent the automaton in each search: block j is used to search P in $T[jn/w..(j+1)n/w - 1]$, for $0 \leq j \leq w - 1$ (we assume w divides n). Since the operations involved in updating the automata are the same across blocks, an update to all w automata can be done with a constant number of single wide word operations. All bit vectors of the precomputed table Y are now again w -bit long, as in the original algorithm. In each step of the search, w entries of Y are read in parallel to each block according to the current character in T in the search in each portion. The pseudocode for this procedure is shown in Algorithm 5. The code assumes $m \leq w$, though it is straightforward to modify it for the $m > w$ case. The running time of this algorithm is now $O(nm/w^2 + n/w + occ)$, where occ is the number of occurrences found. This is asymptotically faster than the first version above, and it leads to Theorem 4.

Algorithm 5 Parallel Shift-And($T, P, n = |T|, m = |P|, \Sigma$). For technical reasons, assume that $T[n + j] = \$$ for $j = 1, \dots, m - 1$, with $\$ \notin \Sigma$, and that $w \geq \log(n + m)$. In order to report matches at each step in time proportional to the number of matches (and not the number of blocks), we move directly to blocks with matching positions by using a function that for every word of length w returns an array A with the positions of set bits. For example, for $w = 5$ and $x = 01011$, $A = [1, 3, 4]$. We do this by table look up to a table with $(w/2)$ -bit entries, whose space is $O(2^{w/2}w)$ words, which for $w = \log n$ is $O(\sqrt{n} \log n)$.

```

1: {Preprocessing}
2: for each  $\sigma \in \Sigma$  do
3:    $Y[\sigma] \leftarrow 0 \{ |Y[\sigma]| = w \}$ 
4: for  $j = 1$  to  $m$  do
5:    $Y[P[j]] \leftarrow Y[P[j]] \mid (1 \gg (j - 1))$ 
6:  $Y[\$] \leftarrow 0$ 
7:  $V \leftarrow \mathbf{0}$ 
8:  $\text{ONES} \leftarrow \frac{2^{w/2} - 1}{2^{w/2} - 1} \{ \text{ONES}_j = 1 \text{ for all } j \}$ 
9:  $C \leftarrow \text{ONES} \gg (w - 1) \{ C_j = 2^{w-1} \text{ for all } j \}$ 
10: {Search}
11:  $n' \leftarrow n/w$ 
12:  $\text{POSNS} \leftarrow \mathbf{0}$  {current positions in text}
13: for  $j = 0$  to  $w$  do
14:    $\text{POSNS} \leftarrow \text{POSNS} \mid ((jn' + 1) \gg wj)$ 
15: for  $i = 1$  to  $n' + m - 1$  do
16:    $V1 \leftarrow (V \gg 1) \mid \text{ONES}$ 
17:    $V2 \leftarrow \text{POSNS}$ 
18:    $\text{read\_content}(V2, T)$  {load characters in each position ( $V2_j = T[\text{POSNS}_j]$ )}
19:    $\text{read\_content}(V2, Y)$  {lookup masks in array  $Y$  ( $V2_j = Y[T[\text{POSNS}_j]]$ )}
20:    $V \leftarrow V1 \& V2$ 
21:    $W \leftarrow V \& C$  {check for matches at each block}
22:    $W \leftarrow \text{compress}(W \ll w - 1)$ 
23:    $\text{matches} \leftarrow W_0$  {matches[j] = 1 if there was a match at block j}
24:    $\text{write\_word}(\text{POSNS}, \text{matching\_positions})$  {write all current positions in array matching\_positions}
25:    $A \leftarrow \text{lookup}(\text{matches})$  {position in  $T$  of  $k$ -th matching block is at matching\_positions[A[k]]}
26:   for  $k = 1$  to  $|A|$  do
27:      $\text{report match at matching\_positions}[A[k]]$ 
28:    $V \leftarrow V \& \sim C$  {clear most significant bit in each block}
29:    $\text{POSNS} \leftarrow \text{POSNS} + \text{ONES}$  {update positions in  $T$  ( $\text{POSNS}_j \leq n + m - 1$  for all  $j$ , thus there is no carry across blocks)}

```

5.2 Boyer-Moore-Horspool

BMH [28] keeps a sliding window of length m over the text T and searches backwards in the window for matching suffixes of both the window and the pattern. More specifically, for a window $T[i..i + m - 1]$, the algorithm checks if $T[i + j - 1] = P[j]$ starting with $j = m$ and decrementing j until either $j = 0$

(there is a match) or a mismatch is found. Either way, the window is then shifted so that $T[i + m - 1]$ is aligned with the last occurrence of this character in P (not counting $P[m]$). The worst case running time of BMH is $O(nm)$ (when the entire window is checked for all window positions) but on average the window can be shifted by more than one character, making the running time $O(n)$ [5]. In the UW-RAM, we can take advantage of the wide word to make several character comparisons in parallel, thus achieving a w -fold speedup over the worst case behaviour of BMH. A recent SIMD-based implementation of BMH using SSE4.2 on Intel i5 and Xeon processors [30] is evidence of the practicality of this approach.

First, we divide each wide word in f -bit fields so that each field contains one character, thus $f = \lceil \log \sigma \rceil$. At each position of the window, we do a field-wise comparison between a wide word containing the characters of the text and one containing the characters of the pattern. We do this simply by subtracting both words. Since we only care if all symbols in the words match, we only need to check if the result is zero, without having to worry about carries crossing fields (and hence we do not need a test bit). We shift the window to the next position if the result is not zero. Note that this check can be done in constant time, and it is quite simple as we do not need to identify where there was a mismatch. Thus in each window we can compare up to w^2/f symbols in parallel, and hence the running time in the worst case becomes $O(mn \log \sigma / w^2 + 1)$. We show the pseudocode in Algorithm 6 which, again, is based on the pseudocode of this algorithm presented in [33, Chapter 2.3.2]. Note that for a given input the distance of the shifts is exactly the same as in the original version of the algorithm, and therefore the average running time remains the same. Note as well that the average running time can be reduced by using each block to search in disjoint parts of the text at the expense of increasing the worst case time to $O(mn \log \sigma / w + 1)$ due to the reduction in the number of characters that can be compared simultaneously.

Theorem 5. *Given T of length n and P of length m over an alphabet of size σ , we can find the occurrences of P in T with a UW-RAM implementation of BMH in $O(mn \log \sigma / w^2 + 1)$ time in the worst-case and $O(n)$ time on average.*

6 Conclusions

We introduced the Ultra-Wide Word architecture and model and showed that several classes of algorithms can be readily implemented in this model to achieve a speedup of $\Omega(\log n)$ over traditional word-RAM algorithms. The examples we describe already show the potential of this model to enable parallel implementations of existing algorithms with speedups comparable to those of multi-core computations. We believe that this architecture could also serve to simplify many existing word-RAM algorithms that in practice do not perform well due to large constant factors. We conjecture as well that this model will lead to new efficient algorithms and data structures that can sidestep existing lower bounds.

Algorithm 6 $\text{BMH}(T, P, n = |T|, m = |P|, \Sigma)$. For simplicity, we assume that w divides $m \log \sigma$. We assume also that T and P are represented with $\log \sigma$ bits per symbol. We still use $T[i]$ to denote one character, which can be easily obtained from the packed representation in constant time (the same applies to the actual address of starting characters of substrings).

```

1: {Preprocessing}
2: for each  $\sigma \in \Sigma$  do
3:    $\text{jump}[\sigma] \leftarrow m$ 
4: for  $j = 1$  to  $m - 1$  do
5:    $\text{jump}[P[j]] \leftarrow m - j$ 
6:  $m' \leftarrow w^2 / \log \sigma$  {characters per wide word}
7: {Search}
8:  $i = 0$ 
9: while  $i \leq n - m$  do
10:   $k \leftarrow m' / m$  {number of window segment}
11:  while  $k > 0$  do
12:     $W \leftarrow T[i + (k - 1)m' + 1..i + km']$  { $W$  contains the substring of  $T$  of  $k$ -th
      window segment}
13:     $V \leftarrow P[(k - 1)m' + 1..km']$  { $V$  contains the substring of  $P$  of  $k$ -th window
      segment}
14:    if  $W - V \neq 0$  then
15:      break
16:    else if  $k = 1$  then
17:      report occurrence at  $i + 1$ 
18:     $k \leftarrow k - 1$ 
19:   $i \leftarrow i + \text{jump}[T[i + m]]$ 

```

References

- [1] AMD: AMD FirePro W9100 Workstation Graphics (Retrieved 11/20/14), http://www.amd.com/Documents/FirePro_W9100_Data_Sheet.pdf
- [2] Andersson, A., Thorup, M.: Dynamic ordered sets with exponential search trees. *J. ACM* 54(3), 13 (2007)
- [3] Arlazarov, V., Dinic, E., Kronrod, M., Faradzev, I.: On economic construction of the transitive closure of a directed graph. *Dokl. Akad. Nauk SSSR* 194, 487–488 (1970), (In Russian). English translation in *Soviet Math. Dokl.*, 11,1209-1210, 1975
- [4] Baeza-Yates, R., Gonnet, G.H.: A new approach to text searching. *Commun. ACM* 35(10), 74–82 (Oct 1992)
- [5] Baeza-Yates, R.A., Régnier, M.: Average running time of the Boyer-Moore-Horspool algorithm. *Theoretical Computer Science* 92(1), 19 – 31 (1992)
- [6] Beame, P., Fich, F.: Optimal bounds for the predecessor problem and related problems. *Journal of Computer and System Sciences* 65, 2002 (2002)
- [7] Bellman, R.: *Dynamic Programming*. Princeton University Press, 1 edn. (1957)
- [8] Bose, P., Chen, E.Y., He, M., Maheshwari, A., Morin, P.: Succinct geometric indexes supporting point location queries. In: *Proc. of SODA*. pp. 635–644 (2009)
- [9] Brodnik, A.: *Searching in Constant Time and Minimum Space*. Ph.D. thesis, University of Waterloo (1995), also available as Technical Report CS-95-41

- [10] Brodnik, A., Carlsson, S., Fredman, M.L., Karlsson, J., Munro, J.I.: Worst case constant time priority queue. *J. of Systems and Software* 78(3), 249 – 256 (2005)
- [11] Brodnik, A., Karlsson, J., Munro, J.I., Nilsson, A.: An $O(1)$ solution to the prefix sum problem on a specialized memory architecture. In: *IFIP TCS*. pp. 103–114 (2006)
- [12] Chan, T.M.: Point location in $o(\log n)$ time, Voronoi diagrams in $o(n \log n)$ time, and other transdichotomous results in computational geometry. In: *Proc. of FOCS*. pp. 333–344 (2006)
- [13] Chan, T.M., Patrascu, M.: Transdichotomous results in computational geometry, i: Point location in sublogarithmic time. *SIAM J. Comput.* 39(2), 703–729 (2009)
- [14] Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C.: *Introduction to Algorithms*. The MIT Press, 2nd edn. (2001)
- [15] Crochemore, M., Iliopoulos, C.S., Pinzon, Y.J., Reid, J.F.: A fast and practical bit-vector algorithm for the longest common subsequence problem. *Inf. Process. Lett.* 80(6), 279–285 (Dec 2001)
- [16] Dantzig, G.B.: Discrete-variable extremum problems. *Operations Research* 5(2), pp. 266–277 (1957)
- [17] Fisher, J.A.: Very long instruction word architectures and the ELI-512. *SIGARCH Comput. Archit. News* 11, 140–150 (June 1983)
- [18] Fredman, M., Saks, M.: The cell probe complexity of dynamic data structures. In: *Proc. of STOC*. pp. 345–354 (1989)
- [19] Fredman, M.L.: The complexity of maintaining an array and computing its partial sums. *J. ACM* 29(1), 250–260 (Jan 1982)
- [20] Fredman, M., Willard, D.: Surpassing the information theoretic bound with fusion trees. *Journal of Computer and System Sciences* 47(3), 424–436 (1993)
- [21] GeForce: GeForce GTX 285 Specifications (Retrieved 11/20/14), <http://www.geforce.com/hardware/desktop-gpus/geforce-gtx-285/specifications>
- [22] Grossi, R., Gupta, A., Vitter, J.: High-order entropy-compressed text indexes. In: *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*. pp. 841–850 (2003)
- [23] Gusfield, D.: *Algorithms on strings, trees, and sequences: computer science and computational biology*. Cambridge University Press, New York, NY, USA (1997)
- [24] Hagerup, T.: Sorting and searching on the word RAM. In: *STACS 98, LNCS*, vol. 1373, pp. 366–398. Springer Berlin / Heidelberg (1998)
- [25] Hampapuram, H., Fredman, M.L.: Optimal biweighted binary trees and the complexity of maintaining partial sums. *SIAM J. Comput.* 28(1), 1–9 (1998)
- [26] Han, Y.: Deterministic sorting in $O(n \log \log n)$ time and linear space. *J. Algorithms* 50, 96–105 (January 2004)
- [27] Han, Y., Thorup, M.: Integer sorting in $O(n \sqrt{\log \log n})$ expected time and linear space. In: *Proceedings of the 43rd Symposium on Foundations of Computer Science*. pp. 135–144. FOCS '02 (2002)
- [28] Horspool, R.N.: Practical fast searching in strings. *Software: Practice and Experience* 10(6), 501–506 (1980)
- [29] Jacobson, G.: Space-efficient static trees and graphs. *Foundations of Computer Science, IEEE Annual Symposium on* pp. 549–554 (1989)
- [30] Ladra, S., Pedreira, O., Duato, J., Brisaboa, N.: Exploiting simd instructions in current processors to improve classical string algorithms. In: *Advances in Databases and Information Systems, LNCS*, vol. 7503, pp. 254–267 (2012)
- [31] Masek, W.J., Paterson, M.: A faster algorithm computing string edit distances. *J. Comput. Syst. Sci.* 20(1), 18–31 (1980)

- [32] Munro, J.I.: Tables. In: FSTTCS. pp. 37–42 (1996)
- [33] Navarro, G., Raffinot, M.: Flexible Pattern Matching in Strings – Practical on-line search algorithms for texts and biological sequences. Cambridge University Press (2002), ISBN 0-521-81307-7. 280 pages.
- [34] Pferschy, U.: Dynamic programming revisited: Improving knapsack algorithms. Computing 63(4), 419–430 (1999)
- [35] Pisinger, D.: Dynamic programming on the word RAM. Algorithmica 35, 128–145 (2003)
- [36] Russell, R.M.: The CRAY-1 computer system. Comm. ACM 21(1), 63–72 (1978)
- [37] Thorup, M.: Combinatorial power in multimedia processors. SIGARCH Comput. Archit. News 31(4), 5–11 (Sep 2003)
- [38] Wikipedia: List of amd graphics processing units (Retrieved 11/20/14), http://en.wikipedia.org/wiki/List_of_AMD_graphics_processing_units
- [39] Wikipedia: List of nvidia graphics processing units (Retrieved 11/20/14), http://en.wikipedia.org/wiki/List_of_Nvidia_graphics_processing_units
- [40] Wu, S., Manber, U.: Fast text searching: allowing errors. Commun. ACM 35(10), 83–91 (Oct 1992)

Appendix

A UW-RAM Subroutines

Comparators Many word-RAM algorithms perform operations on pairs of elements in parallel by packing these elements in *fields* within one word. It is useful to be able to do fieldwise comparisons between two words. Suppose that a word (either regular or wide) is divided in f -bit fields, with each field representing an $(f - 1)$ -bit number. Let G and F be two such words and let F_i and G_i denote the contents of the i -th field in F and G , respectively. Let us assume that we want to identify all F_i such that $F_i \geq G_i$. Fieldwise comparisons can be done by setting the most significant bit of each field in F as a test bit and computing $H = F - G$. The most significant bit of the i -th field in H will be 1 if and only if $F_i \geq G_i$ [24]. Now, if we want to operate only on the values of F that are greater than or equal to their corresponding values in G , we can mask away the rest of the values as follows. We first mask away all but the test bits in H . Then, a mask M with ones in all bits of the relevant fields and zeros everywhere else (including test bits) can be obtained by computing $M = H - (H \ll (f - 1))$. The result of $(M \& F)$ contains then only the values of fields that pass the test [24]. Clearly this operation takes constant time, and it can be easily adapted to other standard comparisons. We shall assume that direct comparisons as well as operations that build on these (such as taking the fieldwise maximum between two words) are available and take constant time [24].